Result Prediction Using K-Nearest Neighbor Algorithm For Student Performance Improvement

Vijay Muralidharan
Department of Information Technology,
Easwari Engineering College,
Chennai, India.
vijaymuralidharan95@gmail.com

Pravien M.
Department of Information Technology,
Easwari Engineering College,
Chennai, India.
mpravien1996@gmail.com

Janani Balaji
Department of Infomation Technology,
Easwari Engineering College,
Chennai, India.
jananibalaji0507@gmail.com

Abstract—**The examination results have become an integral part of every student's life. The educational institution's ranking is greatly influenced by the university results. This paper mainly focuses on the prediction of a student's university result by making use of different attributes. These attributes might be of quantitative and qualitative type. The quantitative attributes used are Internal Assessments, Attendance percentage, Number of On-Duties taken and Overall Assignments completed. The qualitative attributes include Subject feedback, Faculty feedback, and whether the student is a Day Scholar/Hosteller. Here, we make use of k-Nearest Neighbor algorithm (or k-NN for short) against the historical data of students for more accurate prediction of results. In this method all the attributes considered are converted to the same scale. This algorithm makes use of the Euclidean distance formula which is used to find the nearest record. This algorithm predicts better results which help students maximize their academic output.**

Keywords: **University Results, Prediction, Attributes, k-Nearest Neighbor, Euclidean distance formula.**

I. INTRODUCTION

Educational institutions impart knowledge to students and test the imbibed knowledge through exams. These educational institutions are run by both government as well as by private entities teaching engineering, arts, medicine etc…. Examination is an integral part of every educational institution. It also plays a vital role in a student's career.

The exam results form a pivot to judge a student's capacity. The exam results affect the student's future and also influences the ranking of the educational institutions. The ranking of an educational institution solely depends on the results they produce. The result obtained by a student reveals his\her interest and knowledge towards the particular branch of study.

Over the years educational systems have been improving drastically. Result prediction has become a common practice for the improvement of the institution and for the betterment of students. Predicting results by the institutions provides sufficient time to students for improving themselves. This helps in improving the student's academic performance and also the ranking of the educational institution.

Currently the student's academic performance is predicted by considering their assessments and behavioral pattern during class hours. This method of prediction being followed is crude and does not give accurate results. In today's context result prediction plays a major role in not only for ranking of educational institutions but can also be used for much wider applications.

II. RELATED WORKS

Many people have proposed various systems that analyze and predict the behavioral and academic pattern of students. In the year 2014, Naren J [1] proposed a system that specifies the classification techniques for predicting the career options and to predict the violent behavior prevalent among students. A response sheet was used to gather details regarding the background information, reaction of a student when irritated and the interests of a particular student. These parameters were mined to find the corresponding behavioral pattern of a student. However, additional research on student attitudes and learning is needed to predict the patterns efficiently. The response sheet must be created in such a way that it avoids false values and other possible problems. Minimal number of attributes is used which in turn does not give accurate values. Only limited amount of information is suggested for prediction of career options for students.

Authors Nitya Upadhyay, Vinodini Katiyar [2] have proposed a system for predicting the academic performance and the behavior of a student. Various techniques such as Naïve Bayesian Classification, Multilayer Perceptron, J48 and ID3 were used to analyze the attributes. However only a comparative analysis on which technique would give accurate results was specified but not on how they need to be implemented.

III. PROPOSED MODEL

In this section, we are going to present a conceptual model of our system and explain the various aspects of it in a systematic manner.

A. Data Gathering

In order to predict the student results we need to gather large amount of data. Larger amount of data results in more accurate prediction. Hence data gathering plays a vital role. The educational institutions are responsible for collecting and providing these data. Students of the respective institutions have to fill in the necessary details in an excel sheet or fill in a form provided by the institution. These forms include general

IRJECE

information such as name and register number of a student and the necessary information required for prediction.

There are two types of important data required

- Data of past students
- Current student data

The data of past students refers to the collective records of previous batches which are stored in order to predict current student results. Current student data refers to the group of students for whom the prediction process has to be carried out. It is the responsibility of the institution to periodically update and verify the data of the students. This ensures that prediction process yields accurate results.

B. Attribute Selection

There are large amount of student data available which is not being utilized efficiently. Attribute selection is one of the most important steps. It is important to extract meaningful attributes from the gathered data. Only with sufficient and required attributes can the prediction process be implemented and improved. The attributes are classified into:

- Quantitative attributes
- Qualitative attributes

Quantitative data are those which have a specific measure. These are present in the form of numbers. These data are filled in by the faculty based on the student's performance. The attributes used here are Internal Assessments, Attendance percentage, Number of On-Duties taken and Overall Assignments completed. Qualitative data are those which are not present in the form of numbers. It normally defines certain qualities/characteristics. The qualitative data are converted into numerical form by using questionnaires. The questionnaires contain a text box where each student will have to enter a value.

The average value is calculated for the questionnaire based on the value entered by the individual students. These averages are the corresponding numerical value for the respective qualitative attribute of a student. The value given for each option is just a number used for converting qualitative data into quantitative data. The qualitative attributes used are Subject feedback, Faculty feedback, and whether the student is a Day Scholar/Hosteller. The last attribute has a predefined value i.e. Day Scholar is given a value 3, Hosteller as 2, Room staying student as 1. All the attributes selected need to be scaled to a common range. This is to ensure that they can be used in k-Nearest Neighbor algorithm.

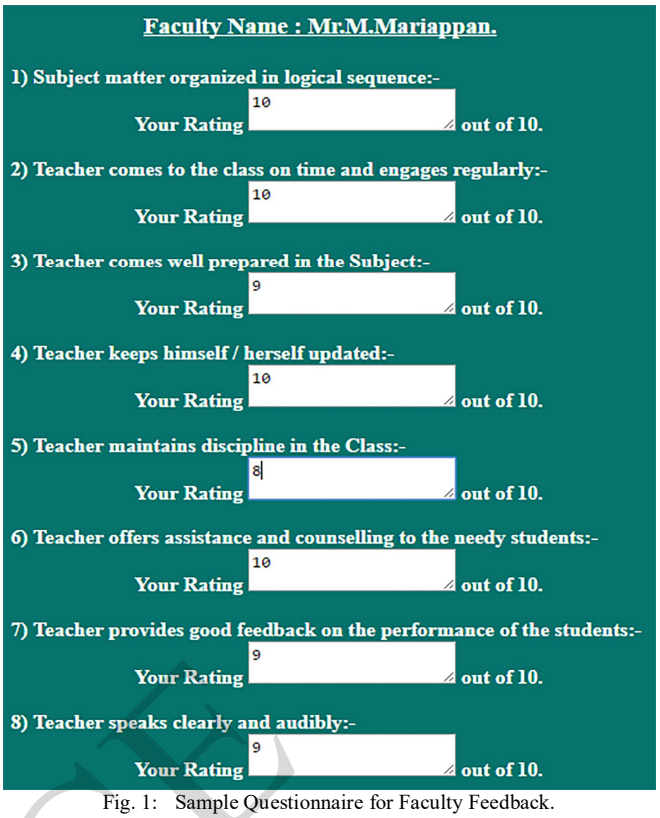


Fig. 1: Sample Questionnaire for Faculty Feedback.

Fig 1 shows a sample form for faculty feedback. The values entered by individual students are taken into consideration while calculating the average. The average for the above sample is calculated to be 9.4. Similarly a questionnaire for subject feedback is used and its average is calculated for each student individually.

C. Implementation

There is numerous numbers of algorithms which are used for prediction process. Some of these algorithms are neural networks, naïve Bayes, k-Nearest Neighbor, decision tree. In this paper we have used k-Nearest Neighbor algorithm for the purpose of prediction.

The main step before prediction process is to check if all the past data of students are available. Only with the help of present and past data of students the prediction process is possible. Once the data is available the required attributes for result prediction are taken into consideration. These attributes are stored in two separate tables. One for the past students data and the next is the current students data for whom the prediction is to be done. The attributes present in the two tables are scaled down to a common range which does not exceed 10. Then with the help of k-Nearest Neighbor algorithm the prediction process is carried out.

The first step for prediction is to randomly set a value for k. This value of k should neither be too large nor too small. This is because k represents the number of nearest neighbor. The value of k cannot be zero or below zero.

The next most important step is the construction of the distance table. This table has to be constructed for each and

IRJECE

every student for whom the prediction process is carried out. The construction of this table is based on the Euclidean's distance formula. The formula is as follows:

$D=\sqrt{(x1-y1)^2 +(x2-y2)^2 +(x3-y3)^2 +\ldots\ldots\ldots + (xn-yn)^2}$.

Where,

D is the distance

x1, x2…xn = the records of current students

y1, y2...yn = the records of past students

Once the distances are calculated using the Euclidean's distance formula they are sorted into ascending order. Now the distance table of each current student consists of distances calculated through their records vs. all the past students records. Since they are sorted it is easier to find the k-nearest neighbor with the help of k value. All the records other than the k-nearest neighbors are removed from the table. Only the records equal to the value of k are left in the distance table.

This shows that the range of the result prediction lies somewhere within the range of these records. After the construction of distance table with only the required records, the past students records are fetched from the past students table. The exam results of these students are checked and the average results of these students are calculated. Once the average has been computed the value is subtracted and added with 0.25. This gives us a range in which the result lies.

IV. SIMULATION

The simulation is carried out using excel for easier representation and understanding. The application of k-Nearest Neighbor algorithm along with Euclidian's distance formula is carried out on a sample data.

Student Register No.	Test 1	Test 2	Test 3	Attendance	OD Taken	Faculty Review	Subject Review	Student Type	GPA
310613205059	5	5.4	6.9	7.8	3.4	7.9	6.7	3	7.2
310613205060	4	8.9	5.8	6.7	5.5	6.9	8.7	2	8.1
310613205061	6	4.6	9.5	8	7.4	8.7	9.1	2	8.5
310613205062	7.8	7.3	6.6	7.4	8.2	6.7	5.6	2	6.4
310613205063	4.4	8.8	7.4	4.5	9	7.7	6.3	3	5.5
310613205064	5.6	9.2	8.5	6	6.6	9.2	3.4	3	7.7
310613205065	9.2	7.8	9.2	5	7.4	6.7	8.2	1	6.5
310613205066	8.8	9.3	5.5	7.6	8.5	8.5	6.7	1	7

Fig. 2: Past Students Table

Student Register No	Test 1	Test 2	Test 3	Attendance	OD Taken	Faculty Review	Subject Review	Student Type	GPA
310614205065	7.3	6.5	5.5	8.7	5.6	9.4	6.6	3	?
310614205066	6.7	7.7	6.3	7.2	8.3	6.7	8.3	2	?
310614205067	6.4	8.6	6.5	5.5	7.7	9.6	7.5	3	?
310614205068	7.4	9.2	8.6	9.2	5.6	8.7	9.2	2	?
310614205069	7.8	7.8	7.4	6.9	5.6	9.4	8.2	2	?
310614205070	9.3	6.5	6.8	9.3	6.7	8.4	8	2	?
310614205071	6.8	7.5	8.7	5.6	7.8	6.2	6.9	3	?
310614205072	9.4	7.3	6.4	8.5	9.1	6.7	5.9	1	?

Fig. 3: New Students Table

Fig 2 consists of sample data for past students record which consists of all the required data along with the final result with which new results are going to be predicted. Similarly Fig 3 consists of sample data of new students. It contains all the attributes required to predict its final result. The student type in both the tables indicates if the student is a day scholar\hosteller.

Past Student Register No.	Distance	Order
310613205059	4.05	1
310613205060	5.69	4
310613205061	5.72	5
310613205062	4.45	2
310613205063	7.03	7
310613205064	6.14	6
310613205065	7.05	8
310613205066	4.95	3

Fig. 4: Distance Table

Let k=3, this tells us that the first three records have to be extracted after the distance has been computed.

The distance is computed using the Euclidean's distance formula. The calculation for student register number 310613205065 against one record is shown:

$D=\sqrt{(7.3-5)^2 + (6.5-5.4)^2 + (5.5-6.9)^2 + (8.7-7.8)^2 + (5.6-3.4)^2 + (9.4-7.9)^2 + (6.6-6.7)^2 + (3-3)^2}$.

D=4.05

Similarly all the distances can be calculated for all the students. The distance table for the above student is generated as in Fig 4. Fig 5 shows the distance table for the student 310613205065 in ascending order.

Now we extract the first three records as the specified value of k is 3. Taking the average of the results of these three

IRJECE

records we get 6.87. Therefore we can compute the range to be 6.62 to 7.12. This is done by subtracting and adding the average value with 0.25.

Past Student Register No.	Distance	Order
310613205059	4.05	1
310613205062	4.45	2
310613205066	4.95	3
310613205060	5.69	4
310613205061	5.72	5
310613205064	6.14	6
310613205063	7.03	7
310613205065	7.05	8

Fig. 5: Sorted Distance Table

Register No.	Prediction Result	Result Range From	Result Range To	Original Result	Within Range
310614205065	6.87	6.62	7.12	6.66	YES
310614205066	7.17	6.92	7.42	6.95	YES
310614205067	6.30	6.05	6.55	7.00	NO
310614205068	7.87	7.62	8.12	8.10	YES
310614205069	6.30	6.05	6.55	6.80	NO
310614205070	7.53	7.28	7.78	7.50	YES

Fig. 6: Final Predicted Table

Similarly all the results for all the students were predicted. Fig 6 shows the final table of predicted results of students and was verified after their results were published.

V. CONCLUSION

The proposed system provides insights of how the student results can be predicted. It showcases the efficient result prediction strategies used under k-Nearest Neighbor algorithm. This when implemented on a large scale by various institutions will yield fruitful results which in turn will maximize the prediction rate and improve the individual academic results of an educational institution, thereby increasing their ranking.

he proposed system gives the institution the time to progress the students in their areas of improvement. Though there have been different classification algorithms predicting the results of students and yet no algorithm has been applied on real time educational datasets. In this work we have computed the results of some students belonging to an educational institution and showed results.

VI. FUTURE SCOPE

In this paper we consider some of the most common attributes from the student's data. A lot of other factors can be taken into consideration for better prediction purposes is left as future scope of this paper. This system can be implemented in a wider scale if a common database is used by all colleges.

With larger amount of data big data analysis can be used for faster prediction of results. It can also be applied for predicting the results of individual subjects which in turn yields more accurate results. This system can also be implemented in other domains such as health sectors. In health sectors symptoms of a patient when considered as attributes can be used for disease predictions.

REFERENCES

[1] Naren J, "Application of Data Mining in Educational Database for Predicting Behavioral Patterns of the Students" IJCSIT, Vol.5 No.03 2014.

[2] Nitya Upadhyay, Vinodini Katiyar, "A Survey on the International Journal of Computer Applications Technology and Research, Vol.3 No.11, 2014

[3] U. bin Mat, N. Buniyamin, P. M. Arsad, R. Kassim, An overview of using academic analytics to predict and improve students' achievement:A proposed proactive intelligent intervention, in: Engineering Education (ICEED), 2013 IEEE 5th Conference on, IEEE, 2013, pp. 126–130.

[4] Shreenath Acharya, Madhu N,"Discovery of students' academic patterns using data mining techniques" IJCSE,Vol. 4 No. 06 June 2012.

[5] T. Mishra, D. Kumar, S. Gupta, "Mining students' data for prediction performance", in: Proceedings of the 2014 Fourth International Conference on Advanced Computing & Communication Technologies, ACCT '14, IEEE Computer Society, Washington, DC, USA, 2014, pp. 255–262.

[6] A.R.Chitupe, S.A.Joshi, 'Classification of ECG Data for Predictive Analysis to Assist in Medical Decisions': IJCSNS International Journal of Computer Science and Network Security, Vol.15 No.10, October 2015.

[7] Sadegh Bafandeh Imandoust And Mohammad Bolandraftar, 'Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events:Theoretical Background': S B Imandoust et al. Int. Journal of Engineering Research and Applications Vol. 3, Issue 5, Sep-Oct 2013, pp.605-610.

[8] Diego Garca-Saiz and Marta Zorrilla, "Comparing classification methods for predicting distance students' performance":MLR: Workshop and Conference Proceedings 17 (2011) 26-32 2nd Workshop on Applications of Pattern Analysis.

[9] G.Narasinga Rao, Srinivasan Nagaraj, "A Study on the Prediction of Student's Performance by applying straight-line regression analysis using the method of least squares":G.Narasinga Rao et al. / International Journal of Computer Science Engineering (IJCSE).

[10] Vaneet Kumar, Dr. Vinod Sharma, "Student's Examination Result Mining: A Predictive Approach" International Journal of Scientific & Engineering Research, Vol.3, No.11 November 2014.

IRJECE